Active Multisensory Perception and LearnIng For InteractivE Robots

Mathieu Lefort[1], Jean-Charles Quinton[2], Marie Avillac[3] and Adrien Techer[1,3]

***Abstract*— The AMPLIFIER (Active Multisensory Perception and LearnIng For InteractivE Robots) project (2018-2022) will study how multisensory fusion and active perception can influence each other during the developmental sensori-motor loop of an autonomous agent. Psychophysics experiments will provide insights on how active perception may influence multisensory fusion in human. Using neural fields, a multi-scale computational neuroscience paradigm, we want to model the behavioral observations in order to transfer and to extend the extracted functional properties to social robots. Especially, we target to provide more natural interactions with humans by allowing the robot to have a better understanding and more appropriate contextual reactions to its environment.**

I. CONTEXT AND OBJECTIVES OF THE PROJECT

An embodied agent senses its environment and the consequences of its performed actions through various channels. They provides many kinds of information related to different perceptual dimensions and environmental conditions to the agent. Thus, to obtain a consistent perception of its surrounding world, the agent has to merge these data. This multimodal merging plays a key role in human perception by improving for instance reaction times, detection thresholds or the learning of stimuli (see [1] for a review). Some sensors may intrinsically have a better performance at some tasks (e.g. a camera has a higher spatial resolution than a microphone). However, the relevance of each sensor in the perception also depends on the current stimuli (a blurred visual stimulus may in some cases be less precise than an auditory localization). So, to generate a multisensory perception, the agent has to solve autonomously two intertwined problems:

- Identification: what stimuli are related to the same environmental event?
- Fusion: what is the relevance of each of these stimuli in this event perception?

The identification problem is related to the detection and learning of previously encountered spatio-temporal co-occurrences. Perception may even be defined as experiencing sensori-motor contingencies [2]. Piaget, by studying the assimilation-accommodation process (i.e. adapting the perception to the environment model while adapting the model to the stimuli) during infant's development, showed that sensori-motor behavioral abilities are appearing in some incremental way [3]. More precisely on the multimodal identification ability, it seems to quickly appear during the development, as also observed at the neuronal level [4]. In machine learning, the study of multisensory representation learning is currently an emerging active field, mainly focusing on image matching with written or oral word/sentence (see [5] for a review).

Concerning the fusion problem, some psychophysics experiments showed that humans are able to fuse information in a Bayes optimal way [6]. This means that the weight of each modality is proportional to its precision, i.e. the inverse of its perceived variance. This Bayesian behavioral model can also be extended in a hierarchical architecture to take into account the identification of the relevant modalities [7]. However, little is known on how an agent can autonomously weight the modality in the perception. It seems that this weighting is updated online by getting more information from the environment, as it decreases progressively when no more cue is provided, similarly to a Kalman filter [8].

Sensori-motor theories [9], which are based on evidences from neuroscience, psychophysics and philosophy, claim that acting plays an important role in this process. This was studied through active perception and predictive coding, yet mainly in monomodal contexts. For instance, saccades can be used to more accurately locate targets [10], with fixations bringing targets in the foveal region of the retina, thus refining the sampled information. Friston develops these ideas in a probabilistic implementation of predictive coding [11] that relies on the free energy principle, which is tightly connected to surprise minimization. This minimization can be achieved by adapting the model (adjusting predictions) or by acting to change the situation (consequently adjusting the stimulation).

Based on this body of evidences, we want to study in this project whether and how active perception plays a role in the multisensory fusion and reciprocally how multiple modalities can improve the selection of action, e.g. related with perceptual attention.

II. CONTENTS OF THE PROJECT

A. Psychophysics experimental protocol

Our experiments are based on the setup used to test the ventriloquist effect [12], where human participants must locate stimuli defined by audio and visual information, whose spatial localization can be either congruent or incongruent. The paradigm allows to test the weighting of modalities in the localization process. To study the influence of active perception on multisensory fusion, we will test whether the

*This project is funded by the Auvergne Rhône-Alpes Region.

[1] LIRIS, SMA team, CNRS UMR 5205, Claude Bernard Lyon 1 University, Lyon University, F-69000, LYON, France name.surname@univ-lyon1.fr

[2] Laboratoire Jean Kuntzmann, ProbaStat / SVH team, CNRS UMR 5224, Université Grenoble Alpes, F-38401, GRENOBLE, France quintonj@univ-grenoble-alpes.fr

[3] Lyon Neuroscience Research Center, Brain Dynamics and Cognition Team, INSERM UMRS 1028, CNRS UMR 5292, Claude Bernard Lyon 1 University, Lyon University, F-69000, LYON, France name.surname@univ-lyon1.fr

ventriloquist effect is moderated by the active perception capabilities of the participant. Operationally, we will compare two experimental conditions, one where saccadic movements are permitted, and one where they are not (relying on eye-tracking techniques). While manipulating the spatial discrepancy between the modalities and other independent variables qualifying the stimuli (e.g., eccenticity, presentation time), we will measure their influence on the visual-auditory interactions. Dependent variables include the explicit location reported by the participant (e.g. by pointing), the location of the visual stimulus on the retina (post-saccade) and the detection threshold. Depending on the obtained results, this experience will be extended to a developmental perspective, by comparing results on various age populations, and to other modalities, especially visio-tactile integration.

B. Computational neuroscience modeling

The neural field paradigm [13] is based on a neuronal modeling with a mean field approach while also explaining some saccadic phenomenons at a behavioral scale [14]. By providing decentralized spatial competition/fusion, it can be combined with the learning of topologically aligned multisensory representations to obtain identification of relevant modalities in perception [15]. We want to extend this last architecture by introducing fusion capabilities in neural fields.

In the superior colliculus, a multisensory subcortical structure related to saccadic movement, visual representations rely on a log-polar encoding with a magnified central area [16]. Thus, central visual stimuli should induce wider neuronal activation than peripheral ones, that should lead to a higher influence of this stimulus in the neural field. Then, following a saccade bringing the target in the central visual field, the visual stimulation should have an higher influence in the multisensory perception. This dynamics will be modeled and tested against the collected psychophysics data.

C. Social robots interactions

Grounding on previous data and models, we will apply our approach to multimodal perception in a social robotics context[1]. When a robot interacts with humans in a crowded environment, it must be able to precisely locate persons and objects based on several cues (e.g. voice or sound, shape or movement). To succeed in correctly interacting with humans for extended periods of time while sticking to predefined scenarios, the robot must keep the attention of the individuals while reaching specific configurations, themselves usually defined by a multisensory conjunction of features.

For this purpose, we will combine multimodal integration with a form of predictive coding. Neural field models can indeed encode predictions, at both the continuous sensorimotor level [14] and the more abstract discrete level [17]. Additional flexibility can be added by dynamically chaining predictions (i.e. performing inference in a continuous or discrete space), for instance relying on a D* algorithm [18]. Predictive multimodal primitives will thus be combined to generate more complex behaviors, and the system will simultaneously adjust primitives parameters and select action to minimize prediction error and uncertainty [11]. In a later step, to learn these primitives in an unsupervised way, we will adopt a constructivist and developmental approach, abstracting from the flow of sensorimotor signals and building on earlier results [19], [20].

[1]in collaboration with Hoomano (`hoomano.com`)

ACKNOWLEDGMENT

The authors want to thank the CNRS' Interdisciplinary Mission which funded the APF[2] (Active Perception For Autonomous Predictive Fusion) project that bootstraps the AMPLIFIER one by implementing the preliminary studies.

REFERENCES

[1] R. Welch and D. Warren, "Handbook of perception and human performance," 1986.

[2] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioral and brain sciences*, vol. 24, no. 5, pp. 939–973, 2001.

[3] J. Piaget, *La naissance de l'intelligence chez l'enfant*. Delachaux et Niestlé Neuchatel, Parés, 1948.

[4] M. T. Wallace and B. E. Stein, "Development of multisensory neurons and multisensory integration in cat superior colliculus," *Journal of Neuroscience*, vol. 17, no. 7, pp. 2429–2444, 1997.

[5] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE transactions on big data*, vol. 1, no. 1, pp. 16–34, 2015.

[6] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[7] K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams, "Causal inference in multisensory perception," *PLoS one*, vol. 2, no. 9, p. e943, 2007.

[8] J. B. Smeets, J. J. van den Dobbelsteen, D. D. de Grave, R. J. van Beers, and E. Brenner, "Sensory integration does not lead to sensory calibration," *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, pp. 18 781–18 786, 2006.

[9] M. Mossio and D. Taraborelli, "Action-dependent perceptual invariants: From ecological to sensorimotor approaches," *Consciousness and cognition*, vol. 17, no. 4, pp. 1324–1340, 2008.

[10] H.-k. Ko, M. Poletti, and M. Rucci, "Microsaccades precisely relocate gaze in a high visual acuity task," *Nature neuroscience*, vol. 13, no. 12, pp. 1549–1553, 2010.

[11] K. Friston, J. Mattout, and J. Kilner, "Action understanding and active inference," *Biological cybernetics*, vol. 104, no. 1, pp. 137–160, 2011.

[12] D. Alais and D. Burr, "The ventriloquist effect results from near-optimal bimodal integration," *Current biology*, vol. 14, no. 3, pp. 257–262, 2004.

[13] S.-i. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," *Biological cybernetics*, vol. 27, no. 2, pp. 77–87, 1977.

[14] J.-C. Quinton and G. L., "A unified neural field model of the dynamics of goal-directed eye movements," *Connection Science*, in press.

[15] M. Lefort, Y. Boniface, and B. Girau, "Somma: Cortically inspired paradigms for multimodal processing," in *IJCNN*. IEEE, 2013, pp. 1–8.

[16] W. Taouali, L. Goffart, F. Alexandre, and N. P. Rougier, "A parsimonious computational model of visual target position encoding in the superior colliculus," *Biological cybernetics*, vol. 109, no. 4-5, pp. 549–559, 2015.

[17] Y. Sandamirskaya and G. Schöner, "An embodied account of serial order: How instabilities drive sequence generation," *Neural Networks*, vol. 23, no. 10, pp. 1164–1179, 2010.

[18] J.-C. Quinton and S. Lengagne, "Multilevel planning and control for robotic manipulation," in *ICSC*, 2015, p. 98.

[19] M. Lefort, T. Kopinski, and A. Gepperth, "Multimodal space representation driven by self-evaluation of predictability," in *ICDL-EPIROB*. IEEE, 2014, pp. 319–324.

[20] S. Mazac, F. Armetta, and S. Hassas, "On bootstrapping sensorimotor patterns for a constructivist learning system in continuous environments," in *International Conference on the Synthesis and Simulation of Living Systems (ALIFE)*, 2014.